Análisis de datos geológicos

Geological data analysis

Merinero Palomares, Raul; Ortega, Lorena; Piña, Rubén; Lunar, Rosario

Dpto. de Cristalografía y Mineralogía de la Universidad Complutense de Madrid, C/José Antonio Novais 2. 28040, Madrid (España). E-mail: rmeriner@geo.ucm.es

Resumen

Un correcto análisis de los datos adquiridos durante la investigación geológica es fundamental tanto para la interpretación de los mismos como para su posterior comunicación a la comunidad científica. Por ese motivo el aprendizaje de una metodología de análisis de datos debe formar parte de la preparación de todo geólogo a lo largo de su formación. En este trabajo se desarrolla una metodología de análisis de datos como parte principal de dicha preparación. Las técnicas propuestas no requieren conocimientos avanzados de estadística por parte del geólogo y tienen como principal objetivo el establecimiento de relaciones entre datos y la elaboración de hipótesis sobre las mismas que deberán ser formalizadas en una fase final de tratamiento estadístico aplicado.

Abstract

An accurate data analysis on geological research is basic for data interpretation and the subsequent communication to the scientific community. For this reason, the learning of a methodology of data analysis must be part of the teaching of geologists along their scientific formation. In this work we develop a methodology of data analysis as the main element of that preparation. The suggested techniques do not require advanced statistical knowledge by the geologist and have as main purpose the establishment of data relations and the

building of hypothesis with them that will be formalized in a final stage of applied statistic.

Palabras clave:

Representación, interpretación, gráficos, análisis, datos

Keywords:

Presentation, interpretation, diagrams, data analysis.

1. Introducción

El Análisis de Datos es un conjunto de técnicas y métodos estadísticos que aplicados de forma global y sistemática a unos datos permiten obtener conclusiones sobre ellos mismos y sobre la población o poblaciones de la que proceden. Todo análisis de datos debe comprender al menos las siguientes fases:

1. Modelización
2. Adquisición y depuración
3. Descripción y representación
4. Establecimiento de relaciones e hipótesis de trabajo
5. Formalización estadística

Las cuatro primeras fases pueden ser realizadas por cualquier geólogo con una preparación adecuada en análisis de datos cuyo desarrollo metodológico es el cometido de este trabajo. El último punto lo podrá realizar el geólogo con conocimientos en estadística aplicada, aunque si la complejidad es alta tendrá que resolverse por especialistas en esta materia.

Algunas de estas fases forman parte, de forma parcial o completa, de los programas de asignaturas de Matemáticas y/o Estadística que se imparten

preferentemente en el primer curso de Grado de Geología. Estas asignaturas tienen en común el enfoque formal Matemático como demuestra el hecho de ser impartidas por docentes de departamentos de Matemática Aplicada, Estadística o Investigación Operativa. Por ese motivo se basan en el aprendizaje de conceptos eminentemente matemáticos como la combinatoria; el cálculo y las distribuciones de probabilidades; las variables y la inferencia estadística. Solo en algunos casos se complementa la formación matemática con el manejo de herramientas informáticas (p.e. Tratamiento estadístico e informático de datos geológicos en el segundo curso de Grado de Geología en la Universidad de Zaragoza). El enfoque del análisis de datos como metodología de investigación no se plantea durante la formación de graduados ni tampoco en la formación de posgrado. La realización de un análisis de datos incompleto tiene dos problemas asociados. El primero es que en las publicaciones de las investigaciones geológicas no se incluyen descripciones y análisis de datos correctos dificultando el seguimiento de todo el proceso de adquisición y descripción de datos (apartados siempre presentes en estos artículos) y por lo tanto la discusión y posterior conclusiones del trabajo de investigación. El segundo que el geólogo investigador no explota toda la información contenida en los datos adquiridos en muchas ocasiones con un coste y esfuerzo muy elevado.

En este trabajo se propone el contenido y desarrollo de una metodología de análisis de datos basada en la aplicación de técnicas eminentemente prácticas de tratamiento y representación de datos que deberá complementarse con la utilización de herramientas informáticas para la realización del análisis de datos. Se ha incluido un apartado final con algunas de las herramientas

informáticas recomendadas. Los autores proponen que la metodología de análisis de datos debería formar parte de seminarios eminentemente prácticos enmarcados en cursos de posgrado (doctorado o máster).

2. Modelización y adquisición de datos

El punto fundamental de todo análisis de datos es definir de manera clara y concisa qué es lo que queremos hacer. Esto implica entre otras cosas identificar la población de estudio y definir toda la información en forma de datos que vamos a manejar de la misma. Los principales datos geológicos serán medidas de magnitudes físicas o químicas (p.e. datos morfométricos, composiciones químicas, parámetros físicos, cartográficos), datos calculados a partir de dichas medidas y datos obtenidos al clasificar muestras.

Modelización: consiste en asociar variables a cada uno de los datos que van a ser adquiridos o calculados. Las variables pueden ser cuantitativas cuando representan datos numéricos y que por lo tanto toman cualquier valor dentro de un intervalo; y cualitativas para los datos categóricos que solo toman 2 o más valores dentro de un conjunto limitado de opciones denominadas modalidades.

Para las variables cuantitativas es necesario definir las siguientes propiedades:

- Rango o intervalo de variación
- Unidades de medida en función de la magnitud que representan
- Número de decimales representativos

La consulta bibliográfica sobre trabajos anteriores es fundamental para asignar valores a las propiedades de las variables cuantitativas.

Para las variables cualitativas solo habrá que definir todos los posibles valores o modalidades que puedan tomar. Es importante asignar un valor numérico a

cada modalidad para poder representar relaciones de orden. Por ejemplo, podemos asignar valores numéricos para expresar el grosor mediante 1, 2, 3, 4 y 5 representando 1 un menor grosor que 3 y este que 5; y a su vez a cada número una etiqueta: muy delgado, delgado, normal, grueso y muy grueso.

Adquisición de datos: es necesario tener en cuenta y definir los siguientes aspectos:

- Errores cometidos durante el proceso de medida
    - Errores sistemáticos. Constantes, controlables y por lo tanto medibles
    - Errores accidentales. Inevitables y no controlables y necesitan del uso de técnicas estadísticas para su manipulación
- Representatividad de los datos adquiridos. El azar no es garantía de representatividad. La representatividad de los datos respecto de la población que pertenecen es requisito para la validez de la investigación
- Precisión y sensibilidad de los aparatos de medida. Se especifica como parte de la descripción del instrumental de medida utilizado. Relacionado con la modelización de los datos y en concreto con las unidades y posiciones decimales
- Número de medidas necesarias. Las medidas necesarias dependen de la variación y de la máxima diferencia esperadas entre el valor medido y el valor real. La consulta bibliográfica permite estimar esta variación

3. Descripción y depuración

Los objetivos de esta fase son el conocimiento de los datos adquiridos y calculados, su descripción y su posterior depuración para detectar valores

anómalos y/o incorrectos. Aunque la descripción de los datos lleva asociada una serie de descriptores numéricos cuya especificación es fundamental en todo trabajo de investigación, como parte de la metodología de análisis de datos se propone la utilización de métodos gráficos como potentes herramientas de descripción de datos. Los trabajos de Cleveland (1993), (1994) y Chambers et al. (1983) se proponen como textos de referencia en la descripción gráfica de datos.

Las variables cualitativas se describen mediante frecuencias absolutas y relativas, acumuladas o no. La representación numérica de las variables cualitativas consiste en tablas en las que se indican las frecuencias de cada una de las modalidades. La representación gráfica visual se realiza en forma de diagramas de sectores y de rectángulos (Cleveland, 1985) estableciendo una correspondencia entre las frecuencias de cada modalidad y el área de los sectores o los rectángulos (ver Figura 1). Las comparaciones entre variables cualitativas se realizarán mejor mediante representaciones gráficas.

Figura 1. Los diagramas de sectores permiten representar porcentajes y además realizar comparaciones, como por ejemplo en este caso con el porcentaje de minerales encontrados en dos muestras K137 (no alterada) y K142 (alterada). Datos de Malitch et al. (2001).

Las variables cuantitativas se describen mediante medidas de posición (media aritmética, mediana, moda y cuantiles), dispersión (rango, error estándar de la media; recorrido intercuartil, varianza, cuasivarianza, desviación y cuasidesviación típica) y simetría (coeficiente de simetría de Pearson). El valor de las mismas junto con el número total de datos debe incluirse como resultado de todo análisis de datos. La representación gráfica de los datos cuantitativos se realiza mediante histogramas (ver Figura 2), diagramas de tallos y hojas y

gráficos de cajas (Chambers et al., 1983). Los métodos gráficos, además de ofrecer una representación rápida y concisa de las medidas de posición, dispersión y simetría, permiten depurar rápidamente los datos y efectuar un análisis básico de normalidad e igualdad de varianzas. Las técnicas estadísticas más utilizadas (punto 5 del análisis de datos) solo son válidas cuando los datos se han obtenido de forma independiente (fácilmente asumible), siguen una distribución estadística normal (por lo tanto simétrica) y existe homogeneidad de varianzas (los límites de variación son similares entre variables comparadas). Ambos supuestos pueden contrastarse fácilmente en las representaciones gráficas de variables cuantitativas aunque la formalidad se asignará mediante técnicas estadísticas. Una alternativa al cumplimiento de estos requisitos es el uso de medidas posición robustas (mediana, medias recortadas o Windsorizadas) y métodos estadísticos robustos en el punto 5 del análisis de datos, obteniéndose conclusiones formales sin necesidad de normalidad e igualdad de varianza (García-Pérez, 2005).

Figura 2. Los histogramas permiten representar gran cantidad de datos y realizar un análisis de los mismos de manera rápida. En este ejemplo partiendo de 1121 análisis de núcleos inalterados de cromitas se han representado los histogramas de Al_2O_3 y el número de Cromo (Datos de Gervilla et al., 2005). En el histograma de Al_2O_3 se observa la existencia de varias modas una situada en torno a 15 (cromitas ricas en Cr) y la segunda en torno a 30 (cromitas pobres en Cr). En el histograma de Cr# también se observa la división en dos poblaciones, la primera en torno a 0,47 y la segunda en torno a 0,7.

4. Establecimiento de relaciones e hipótesis de trabajo

El objetivo fundamental del análisis de datos, además de estimar los parámetros de posición y dispersión de la población de estudio, es realizar comparaciones y establecer relaciones entre variables de manera que alguna de las variables explique la variación de otras. Las relaciones entre variables se

establecerán en forma de hipótesis de trabajo que deberán ser formalizadas en la fase 5 de aplicación de técnicas estadísticas.

Según los tipos de variables implicadas tenemos tres tipos de relaciones.

4.1. Relación entre variables cuantitativas.

Un primer objetivo al comparar variables cuantitativas es relacionarlas mediante una función, siendo la relación lineal la más usada y fácil de entender. En estadística se conoce como regresión lineal y tiene varios pasos:

- Decidir si existe o no relación lineal entre las variables (correlación)
- Cuantificar la fuerza de dicha relación mediante el coeficiente de correlación
- Obtener los coeficientes de la pendiente de la recta y el corte con el eje de ordenadas (regresión lineal)

La representación gráfica de los valores en un diagrama de dispersión (ver Figura 3) es la manera más rápida de ver si existe relación lineal entre las variables (Cleveland, 1985).

Figura 3. Representación gráfica de relaciones entre variables cuantitativas. Izquierda, diagrama de dispersión con recta de regresión de mínimos cuadrados superpuesta y cuantificación de la regresión para datos del volumen encefálico de recién nacidos y adultos de cuatro especies de homínidos (datos de Bermúdez 2010) donde se observa regresión lineal. Derecha, histograma para comparar diámetros de framboides de pirita obtenidos en el laboratorio y observados en la naturaleza (datos de Merinero 2008) y donde se observa una gran similitud entre ambas poblaciones.

Este tipo de relación puede extenderse a relaciones no lineales estudiando el diagrama de dispersión y comparando con los diferentes tipos de funciones matemáticas o aplicando transformaciones a las variables (Box-Cox, 1964).

Un segundo objetivo al comparar variables cuantitativas es realizar un análisis de homogeneidad. Este tipo de análisis se realiza cuando queremos comparar muestras obtenidas en diferentes condiciones (por ejemplo en el laboratorio y

en la naturaleza). La hipótesis de trabajo en este caso es que ambas poblaciones son equivalentes. La comparación gráfica de histogramas de ambas variables es la manera más rápida de ver si existe homogeneidad de muestras (ver Figura 3).

4.2. Relación entre variables cualitativas.

Cuando comparamos variables cualitativas se pueden establecer las siguientes hipótesis

- Existe dependencia entre las variables
- Existe dependencia solo entre algunas de las modalidades

El cálculo fundamental en este tipo de relaciones consiste en obtener las frecuencias del cruce de cada una de las modalidades de las variables cualitativas (número de datos que cumplen pares de modalidades de cada variable), cuyo resultado es una tabla de doble entrada o de contingencia.

Figura 4. Representación gráfica de relaciones entre variables cualitativas. Izquierda, diagrama de rectángulos acumulados para representar la tabla de contingencia entre engrosamiento (3 modalidades y redondez (6 modalidades) de partículas sedimentarias recogidas en distintos tramos de un río. Las diferencias entre la distribución de porcentajes de redondez para cada modalidad de engrosamiento nos dice que existe dependencia entre ambas variables. Derecha, representación en un gráfico de correspondencia de las diferentes modalidades de ambas variables observándose la relación entre cada tipo de engrosamiento y redondez de las partículas (Barrios et al. datos inéditos).

Una primera aproximación para la elaboración de hipótesis es comparar los valores obtenidos en la propia tabla de contingencia. La segunda, más recomendada, es representar las frecuencias de la tabla en gráficos de rectángulos acumulados (ver Figura 4) y comparar las diferencias de frecuencias entre modalidades (mayor diferencia más probable que exista dependencia). Según aumenta el número de modalidades de las variables es más difícil que exista dependencia total entre las variables. La alternativa en

este caso es realizar un análisis de correspondencia que permite establecer relaciones de dependencia entre modalidades y agrupar aquellas que presenten mayor afinidad. El análisis de correspondencia se realiza también de forma gráfica mediante biplots (Gower and Hand 1996; ver Figura 4).

4.3. Relación entre variables cuantitativas y cualitativas

Finalmente es interesante comparar el efecto de las modalidades de una variable cualitativa en los parámetros de posición, dispersión y simetría de una variable cuantitativa (extensible a varias variables cuantitativas con magnitudes y unidades similares). Para ello se divide la variable cuantitativa en tantos subconjuntos como modalidades presenta la variable cuantitativa, obteniendo de esta manera varias nuevas variables cuantitativas que por lo tanto tienen sus propios parámetros de posición, dispersión y simetría.

Las hipótesis de trabajo que se realizan en este tipo de relación son las siguientes:

- Las variables de posición/dispersión de todas/algunas de las modalidades son iguales
- Las variables de posición/dispersión de todas/algunas de las modalidades son mayores/menores entre sí

Aunque las hipótesis son formalizadas de nuevo mediante test estadísticos (análisis de la varianza con comparaciones múltiples) primero será necesario plantearlas según los valores obtenidos de los parámetros de de las nuevas variables. Si dos de esas nuevas variables presentan medias muy distintas no tendrá sentido plantear una hipótesis de trabajo de igualdad de medias.

Los diagramas de cajas permiten realizar de forma rápida una comparación de los parámetros de posición y dispersión de varias variables cuantitativas (McGill et al. 1978; Williamson et al. 1989) (ver Figura 5).

Figura 5. Utilización de gráficos de cajas para comparar el origen de muestras de cromititas recogidas en diferentes yacimientos de Cuba según el número de cromo y el Al_2O_3 (líquido). Datos de Gervilla et al. (2005). A partir de los gráficos es posible realizar fácilmente hipótesis sobre la equivalencia del origen de muestras con distintas procedencias.

5. Herramientas recomendadas para análisis de datos

- Hojas de cálculo. Fáciles de utilizar y con cálculos potentes, su uso es muy recomendable para la representación y manejo de datos. Sus opciones estadísticas son escasas aunque útiles. Gráficos de gran calidad pero limitados (De Levie 2004)
- Paquetes estadísticos comerciales, tipo SPSS. Accesible a todo tipo de usuarios con varios niveles de complejidad. El análisis de datos se puede realizar de forma completa y cada vez incorporan gráficos de mayor calidad. Su elevado precio así como sus funcionalidades avanzadas limitadas son las únicas desventajas (Sweet 2011)
- Software de estadística de libre distribución. En este campo destaca el software libre R que a la ventaja de su coste cero se une la potencia que añaden los usuarios al desarrollar nuevos métodos estadísticos de análisis y representación gráfica de datos. Recomendado para todo tipo de usuarios (Murrel 2005; Chambers 2008; García-Pérez 2010; R Development Core Team 2010).

Agradecimientos

Este trabajo se ha desarrollado dentro del proyecto número 26 aprobado en la Convocatoria de Proyectos de Innovación y Mejora de la Calidad Docente para el curso 2011/2012 de la Universidad Complutense de Madrid.

Bibliografía

Bermúdez de Castro, J.M. (2010). *La evolución del talento*. Editorial Debate, Barcelona.

Box, G. E. P., Cox, D. R. (1964). *An analysis of transformations (with discussion)*. Journal of the Royal Statistical Society B, 26, 211-252.

Chambers, J. M. (2008). *Software for Data Analysis: Programming with R*. Springer Verlag, Nueva York.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Chapman and Hall, Nueva York.

Cleveland, W.S. (1994). *The elements of graphing data*. Hobart Press, Summit, Nueva Jersey.

Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, Summit, Nueva Jersey.

De Levie, R. (2004). *Advanced Excel for scientific data analysis*. Oxford University Press, Nueva York.

García-Pérez, A. (2005). *Métodos avanzados de estadística aplicada: Métodos robustos y de remuestreo*. Editorial UNED. Colección Educación Permanente.

García-Pérez, A. (2010). *Estadística básica con R*. Editorial UNED. Colección Grado.

Gervilla, F., Proenza, J.A., Frei, J.M., González-Jiménez, C.J., Garrido, J.C., Melgajero, A., Meibom, A., Díaz-Martínez, R., Lavaut, W. (2005). *Distribution of platinum-group elements and Os isotopes in chromite ores from Mayarí-Baracoa (eastern Cuba)*. Contributions to Mineralogy and Petrology 150, 589-607.

Gower, J.C. and Hand, D.J. (1996). *Biplots*. Chapman & Hall, Londres.

Malitch, K.N., Melcher, F., Mühlhans, H. (2001). *Palladium and gold mineralization in podiform chromitite at Kraubath, Austria*. Mineralogy and Petrology 73, 247-277.

McGill, R., Tukey, J.W., Larsen, W.A. (1978). *Variations of Box Plots*. The American Statistician, 31, 12-16.

Merinero, R. (2008). *Procesos mineralógicos y geoquímicos en chimeneas submarinas de carbonatos metanógenos del Golfo de Cádiz: biogeomarcadores framboidales de sulfuros y oxihidróxidos de hierro*. Tesis Doctoral. Universidad Complutense de Madrid.

Murrell, P. (2005). *R Graphics*. Chapman & Hall/CRC Press, Boca Ratón, USA.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Viena, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sweet, S.A., Grace-Martin, K. (2011). *Data Analysis with SPSS: A First Course in Applied Statistics*. Prentice Hall, Boston, USA.

Williamson, D.F., Parker, R.A., Kendrick, J.S. (1989). *The box plot: a simple visual method to interpret data*. Annals of Internal Medicine 110, 916-21.